Metagenomic Contigs Binning with Probabilistic k-mers Statistics

Qian J(1)[†] , Comin M(1)

(1) Department of Information Engineering, University of Padova, Padova, Italy

[†] Email: comin@dei.unipd.it

Motivation

Sequencing technologies allow the sequencing of microbial communities directly from the environment without prior culturing.

The assembly of metagenomic reads typically produces only genome fragments, also known as contigs. Taxonomic analysis of microbial communities requires contig clustering, a process referred to as binning, in which contigs are grouped into putative species. The major problems are the lack of taxonomically related genomes in existing reference databases and the uneven abundance ratio of species. Here we present MetaCon a tool for metagenomic contigs binning based on probabilistic k-mers statistics.

Methods

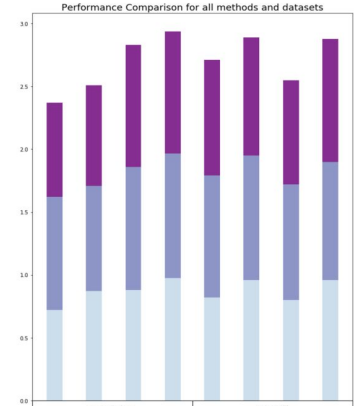
Most binning tools are based on similarity measures between contigs that are built over k-mers frequency distributions. However, when dealing with a similarity measure based on k-mers counts there are two major issues. The first one is that k-mers might have a different probability to appear in different species. The second is that long contigs carry more information than short ones, therefore, the direct comparison between them can be challenging. The first problem has been extensively studied in the field of alignment-free measures. The latter, suggests that short contigs should be treated differently. MetaCon addresses these problems by proposing a two-phases binning algorithm in which each phase process one portion of the input dataset. Let us assume that we have N contigs to group into bins. Following past studies, the composition of contigs (in terms of its constituent k-mers) and the abundance (or coverage) information, that is the average coverage of contigs, are promising features.

Firstly, we construct the feature matrix, where every row corresponds to a single contig that is represented by a feature vector where some features are from the coverage matrix Y and the rest are from the composition matrix Z. The coverage matrix is rescaled by the sum of columns (across the contigs) and then by the sum of row (across samples). As we mentioned before the length of contigs may play an important role in clustering, we suggest to individually deal with the short and long contigs. We split the composition and coverage matrices into two sub-matrices according to the length distribution of contigs, indicated as Zl, Zs, Yl, Ys. In the first phase, we normalize the composition matrix of long contigs by means of probabilistic k-mers statistics. Since contigs are from different species, and therefore the underlying k-mers distributions are different, we compute expectation and variance of k-mers counts for each contig, based on a probabilistic model of k-mers

statistics. Then, k-mers counts of each contig in the matrix Zl are normalized independently of the other contigs, based on the corresponding probabilistic k-mers distribution profile.
Then the feature matrix, composed by Yl and Zl, is feed into k-medoids to cluster long contigs into bins. In the second phase, we assign the short contigs to the closest centroids, the outcome of the first phase, by measuring the shortest L1 distance between the feature vector and the centroids.

Results

We compared MetaCon in terms of precision and recall against other popular binning methods: CONCOCT, MetaBat and MaxBin. We assess the performance based on three datasets also used in other studies (see Figure): two synthetic and a real metagenome. For 'Strain' dataset, the precision of MetaCon is about 97.5%, better than the other three methods; the recall is 95.8%, higher than MaxBin and MetaBat, almost identical with CONCOCT. For the 'Species' dataset, it is challenging to bin the contigs since the number of species (101) is large, MetaCon reaches 99.3% in terms of precision and 94.6% for the recall. Regarding to the real dataset 'Sharon', the results are in line with those of the synthetic datasets. MetaCon outperforms CONCOCT, MetaBat and MaxBin in terms of overall precision, recall (see Figure) and as well as the quality of every single cluster.



	Precision				Recall			
	MaxBin	CONCOCT	MetaBat	MetaCon	MaxBin	CONCOCT	MetaBat	MetaCon
Strain	0.72	0.87	0.88	0.975	0.82	0.96	0.8	0.958
Species	0.9	0.84	0.98	0.99	0.97	0.99	0.92	0.94
Sharon	0.75	0.8	0.97	0.97	0.92	0.94	0.83	0.98

Full paper available here: http://www.dei.unipd.it/ ciompin/papers/bits/metacon2.pdf